Többszavas kifejezések számítógépes kezelése

Oravecz Csaba, Varasdi Károly és Nagy Viktor

MTA Nyelvtudományi Intézet, Budapest
{oravecz,varasdi,nagyv}@nytud.hu

Kivonat Azok a több szóból álló kifejezések, melyek tulajdonságainak egy része nem következik a nyelvtan sztenderd szabályaiból, a nyelvi elemzés valamilyen szintjén egy egységként jelennek meg. A számítógépes nyelvfeldolgozás során ezeket az egységeket képesnek kell lennünk azonosítani, és hozzájuk rendelni a produktív szabályokból nem következő jellemzőiket. Ez a feladat a hazai számítógépes nyelvészetben súlyához képest eddig kevés figyelmet kapott, ezért a dolgozat egy lehetséges azonosító, kinyerő módszer értékelése mellett a többszavas kifejezések kezelésének általános problémáit is tárgyalja.
Kulcsszavak: többszavas kifejezések, szinonimitás, idióma, lexikai idioszinkrázia, helyettesíthetőség

1. Bevezető

A számítógépes nyelvfeldolgozás szinte minden területén, különösen pedig a minél részletesebb („mély") nyelvi elemzést igénylő feladatokban a többszavas kifejezések kezelése az egyik legnagyobb kihívást jelentő probléma. Egészen a legutóbbi időkig még az ilyen irányú nemzetközi kutatást sem tartották elegendő mértékűnek (Sag et al., 2002), magyar nyelven pedig egy viszonylag jól körülírható típussal foglalkozó kezdeti próbálkozások kivételével[1] a probléma kutatása éppen, hogy elkezdődött (Kis et al., 2004; Kis és Ugray, 2003). Fontosnak tartjuk ezért azt, hogy a magyar nyelvvel foglalkozó számítógépes nyelvfeldolgozás szempontjait tekintve elsődlegesnek néhány alapvető kérdést megpróbáljunk összefoglalni és tisztázni. Kiindulásként kísérletet teszünk annak meghatározására, hogy a magyar nyelvben milyen szósorozatokat tekinthetünk *többszavas kifejezésnek* (TSZK), ezeknek milyen típusait érdemes megkülönböztetni, milyen problémát jelent a számítógépes kezelésük, s milyen eszközök és eljárások állhatnak rendelkezésünkre, ha ezt a problémát le szeretnénk küzdeni. Mint minden (kezdeti állapotban lévő, folyó) kutatásban, nagy mértékben támaszkodunk a külföldi szakirodalomra, de egyúttal szeretnénk azt is bemutatni, hogy az eddigi eredmények milyen módon hasznosíthatók, illetve milyen sajátos problémákkal kerülünk szembe a magyar nyelvre irányuló számítógépes nyelvészeti kutatásokban.

[1] A „nyílt tokenosztályok" kezelése (*named entity recognition*). Éppen ezért ezzel a típussal a jelen dolgozat csak az említés szintjén foglalkozik.

A dolgozat a következőképpen épül fel. A 2. rész rövid leírást ad a TSZK-król általában. A 2. rész a TSZK-k típusait tárgyalja, míg a 3. részben számítógépes kezelésük problémáit mutatjuk be. A 4. rész lehetséges nyelvi diagnosztikai eszközöket ír le, majd az 5. rész egy kinyerő/azonosító eljárást értékel, és megmutatja, hogy az alapjául szolgáló hipotézis a magyar nyelvű korpuszban nem érhető tetten. Rövid összefoglalás zárja a dolgozatot a 6. részben.

2. Mi az a TSZK?

Az, hogy valójában mit tekintünk többszavas kifejezésnek, vizsgálható a nyelvleírás, nyelvelmélet (Chomsky, 1980; Pulman, 1993; Nunberg et al., 1994) és a számítógépes nyelvfeldolgozás, illetve ennek részterületei szempontjából is.[2] Mivel még az előbbi is jellemzően hol szűkebb, hol tágabb értelemben vizsgál több szóból álló kifejezéseket, meglehetősen reménytelen próbálkozás egy olyan egzakt, részletes definíciót adni, ami aztán a számítógépes nyelvészetben is minden területen egyformán jól alkalmazható lenne. Legáltalánosabban Sag et al. (2002) és Calzolari et al. (2002) nyomán a következőt lehetne definíciónak tekinteni: számítógépes nyelvészetben többszavas kifejezésnek (TSZK) nevezünk egy olyan idioszinkratikus értelemmel rendelkező szósort[3], ami a nyelvi elemzés valamilyen szintjén egy egységként jelenik meg. Ez aztán az alábbiak közül egy vagy több tulajdonsággal rendelkezhet:

- nem teljes kompozícionalitás, nem teljesen megjósolható jelentés. A kifejezés vagy megjósolhatatlan módon kétértelmű: a szószerinti jelentése mellett van egy átvitt értelme is, ami nem tudható be sem lexikai, sem szintaktikai kétértelműségnek (*húzza a lóbőrt, kiborítja a bilit*), vagy van valamilyen hozzáadott jelentés, ami nem megjósolható az elemek jelentéséből (*kaparós sorsjegy*).
- többé kevésbé rögzített forma. A kifejezés egy folytonos skálán helyezkedik el, melynek egyik végén az egyáltalán nem változtatható kifejezések állnak (*ad hoc*), majd azok, ahol szerepel legalább egy olyan elem, ami nem létezik más környezetben (*hiszi a piszi, kidobja a taccsot*), vagy elemcsoport, ami ilyen (*a füle botját se mozdítja*), másikon a egyes elemek helyettesítését nem toleráló TSZK-k állnak (**vakarós sorsjegy*[4]).
- megsérti a szintaxis szabályait (*Ezt nevezem!* vs. *ezt vminek nevezem*).

A TSZK-k gyakorlatilag a nyelvtan és a lexikon közötti területen helyezkednek el. Ezekből a tulajdonságokból, valamint abból a tényből, hogy a TSZK-k

[2] A bevezetőben említett előzménynélküliség bizonyos mértékig a magyar általános nyelvészeti szakirodalmat tekintve is igaz. Lásd pl. Forgách (1998), aki „mostohán kezeltnek" tekinti a területet.

[3] A *szósor* inherens tulajdonsága, hogy szóhatárt, ami általában szóköz karakter, tartalmaz.

[4] Megjegyzés: a következőkben a '*' jel nem feltétlenül szintaktikai rosszulformáltságot jelöl, hanem azt, hogy egy idiomatikus kifejezés egy adott változatban már nem rendelkezik idiomatikus olvasattal.

száma igen nagy, illetve a fenti jellemzőket tekintve egy a teljesen kompozícionális, produktív szókapcsolatoktól a teljesen idiomatikus és rögzített alakokig tartó kontinuum mentén helyezkednek el, több probléma következik (mind elméleti mind) számítógépes szempontból. Egyrészt nehéz egzakt, általános kategóriarendszer alapján adott típusokba sorolni, illetve egyáltalán azonosítani őket. Az, hogy adott esetben szükség van-e arra, hogy adott kombinációt TSZK-nak tekintsünk, függ a feldolgozás módjától és mélységétől (pl. egy szintaktikailag reguláris TSZK nem releváns, ha a rendszer kimenete csupán szintaktikai elemzés). Másrészt, mivel ezek a szókapcsolatok olyan jellemzőkkel rendelkeznek, melyek nem következnek a nyelvtan sztenderd szabályaiból, az adott TSZK-khoz hozzá kell tudni rendelni ezeket a jellemzőket. Ehhez a nyelvfeldolgozás során természetesen képesnek kell lennünk azonosítani a kifejezéseket, illetőleg a lexikonban történő eltárolásukhoz alkalmas reprezentációs formalizmust kell felhasználnunk. A következő részben javasolt osztályozás után mind a két problémát tárgyaljuk az alábbiakban.

2.1. TSZK típusok — egy lehetséges osztályozás

A TSZK-k osztályozásában sincs sokkal nagyobb konszenzus és általánosan elfogadott sztenderd, mint az azonosításukban. Mindesetre két alapvető szempontot meg tudunk különböztetni, és ennek alapján kétféle kategorizálást is végezhetünk: egy operacionálist és egy grammatikai tulajdonságokon alapulót. Természetesen mindegyik osztályozás jár bizonyos következményekkel a számítógépes feldolgozásban használható módszerek és lehetőségek tekintetében.

Az előbbi szempont szerint triviális módon, a TSZK-ban szereplő elemek egymás után következésének függvényében különböztethetünk meg kétféle típust, **folytonos** (*bakot lőtt*) és **megszakított** (*megkongatja a vészharangot* vs. *a vészharang, amit a tudósok megkongattak*) kifejezéseket. Ugyan a TSZK-ban szereplő elemek sorrendje sokszor kötött (van egy *kanonikus sorrendje* a szavaknak), a magyar nyelvben igen gyakran találhatunk sorrendi variánsokat, akár nem a TSZK-hoz tartozó elemek beékelődésével együtt. Ezzel a problémával mindenképpen szembekerülünk a számítógépes elemzés/feldolgozás során.

A grammatikai jellemzők alapján Bauer (1983), Nunberg et al. (1994) és Sag et al. (2002) terminológiáját és osztályozását felhasználva az alábbi típusokat különíthetjük el:

- **intézményesült kifejezések**. Szintaktikailag és szemantikailag kompozícionálisak, statisztikailag idioszinkratikusak[5], vagyis az adott jelentés hordozóiként sokkal gyakrabban fordulnak elő, mint más hasonló kompozícionális kombinációk, illetve akár blokkolhatják is a jelentés alternatív realizációját[6].
- **funkcióigés kifejezés**. A teljesen lexikalizált alakoktól (*részt vesz, pofon vág*) a „terpeszkedő kifejezésekig" (*javaslatot tesz*).

[5] Talán azt mondhatjuk, hogy a „sztochasztikus kompozícionalitás" feltételének nem felelnek meg.

[6] Pl. *közlekedési lámpa* vs. **forgalmi lámpa*, *közlekedési *világítótest*. Az ilyen módon blokkolt szókapcsolat egyébként *anti-kollokáció* néven ismert (Pearce, 2001).

- **ige + partikula szerkezet.** Ide tartoznak az igekötős igék elváló alakjai (akár ragos névszós igekötővel: *létrehoz, észrevesz*).
- **féligáttetsző idiómák.** Összetevői jól megfeleltethetők az idiomatikus jelentésükben szereplő összetevőknek. Ezt onnan láthatjuk, hogy pl. *bakot lő* idióma *bak* részkifejezése szisztematikusan módosítható olyan kifejezésekkel, amikkel a *hibákat* lehet jellemezni: *nagy/komoly/végzetes/elképesztő bakot lőtt* ill. *nagy/komoly/végzetes/elképesztő hibát követett el.* Ez a szisztematikusság hiányzik a következő típusban szereplő kifejezésekből.
- **homályos idiómák.** Szemantikai transzparencia hiányában ezeknek a kifejezéseknek az összetevői nem vethetők alá módosításoknak az idiomatikus jelentés elveszése nélkül. Ennek következtében ez a típus gyakorlatilag csak inflexiós variánst tűr meg (*felveszi a kesztyűt* vs. *felvette a *politikai kesztyűt*).
- **többszavas tulajdonnevek.**
- **összetett szavak.** Pontosabban a helyesírási szabályok miatt szóközt tartalmazó összetett szavak (*nagy néha, nyitva tartás*).
- **rögzített kifejezések.** Semmilyen módosítást nem engednek meg, csak egyféle alakban léteznek (*így vagy úgy, egytől egyig*).

Ez az osztályozás, amint látni fogjuk, sajnos nem képezhető le közvetlenül a feldolgozó eljárások különböző típusaira.[7]

3. A TSZK-k feldolgozásának problémái

3.1. TSZK-k azonosítása, korpuszból történő kinyerése

A TSZK-k számítógépes elemzésének legelfogadottab módja lexikonban való tárolásuk, és abból történő kiolvasásuk[8]. Ebből következően a legfontosabb feladat korpuszból történő automatikus kigyűjtésük és osztályozásuk, amelynek érdekében ritkán tisztán szimbolikus, előre meghatározott szabályokon alapú (Bourigault, 1996), sztohasztikus (Church és Hanks, 1990; Dunning, 1993; Shimohata et al., 1997), illetve leginkább vegyes rendszereket (Daille, 1996; Heid, 1999) használnak. Az előbbiek alapvető problémája a nyelvspecifikusság, ám a klasszikus statisztikai alapú rendszereknek is van egy komoly hátránya. Ezek ugyanis a

[7] Érdemes itt egy terminológiai kitérőt is tenni és megjegyezni, hogy a gyakran használt *kollokáció* kifejezés alatt sokak leginkább az itt *intézményesült kifejezésnek* nevezett típust értik (McKeown és Radev, 2000). Elterjedt azonban egy jóval átfogóbb értelmezés is, amely szerint kollokáció minden szignifikánsan gyakran előforduló szókapcsolat, vagyis az összes fenti TSZK (Manning és Schütze, 1999), valamint akár az egyéb (nem nyelvi) okok miatt gyakran együtt előforduló teljesen produktív és kompozícionális szókapcsolat is (pl. ⟨*apróhirdetés, felad*⟩) (Sag et al., 2002).

A Nunberg et al. (1994) által bevezetett szemantikai dekomponálhatóság szempontjából szokás egyébként az itt féligáttetszőnek ill. homályosnak nevezett idiómákat dekomponálható ill. nemdekomponálható idiómáknak is nevezni.

[8] Bár van példa futási időben működő, általában szabály alapú TSZK elemző rendszerre is (Oflazer et al., 2004).

TSZK-knak azt az egyik fontos tulajdonságát használják ki, hogy elemeik általában gyakrabban fordulnak elő együtt, mint egyéb önkényes szókombinációk, és ennek az együtt előfordulásnak az erősségét számszerűsítik valamilyen *asszociációs mérték* (AM) segítségével. Csupán ennek a mértéknek a használata azonban rendkívül zajos eredményre vezet, különösen, ha a gyakorisági alapú mutató megbízhatóságának növelése és az osztályozás finomítása érdekében minél nagyobb mennyiségű korpuszt használunk. Ez a probléma szabad szórend esetén még élesebben jelentkezik (Kaalep és Muischnek, 2002).

Egy lehetséges megoldás egyrészt a feladat leszűkítése (vagyis a feladat nem általában TSZK-k keresése, hanem valamelyik jól meghatározott összetevőkből álló és szerkezetű részosztályé[9]), ehhez azonban szükséges a felhasznált korpusz minél részletesebb nyelvi annotációja, és ennek az annotációnak a legalaposabb kihasználása. Kérdés azonban, hogy honnan származik és egyáltalán rendelkezésre áll-e ez az annotáció?

Természetesen az ideális eset emberi tudás felhasználását nem igénylő nemfelügyelt tanuló eljárás(ok) alkalmazása lenne, amely nyers korpuszból képes lenne adott típusú TSZK azonosítására, és ugyan történtek kísérletek ebben az irányban (Schone és Jurafsky, 2001), biztató eredményt nemigen hoztak. Jelenleg tehát kénytelenek vagyunk opportunista megközelítést választani: mivel nincs egyedül célravezető, minden típusú TSZK kinyerési feladatra alkalmas módszer (Krenn és Evert, 2001), szűkítsük a szóba jöhető jelöltek körét egy jól meghatározott típusra, és használjunk fel minden lehetséges nyelvi erőforrást az adott típushoz igazított kinyerési módszerhez.

Ezt a megközelítést magyar nyelvre több szempont is indokolja. Ha a szórendi változatosságot legalább bizonyos mértékig figyelembe vevő nem minimális számú (2-3) szóból álló szókapcsolatok tetszőleges sorozataiból képezünk jelöltlistát, a nagy számú, nagy variábilitással bíró elem miatt nagyobb korpusz használata esetén implementációs, hatékonysági problémákba ütközhetünk, ezért célravezető a jelöltlista típusos szűkítése. Ehhez persze a korpusz minimális (POS) annotációjára legalább szükség van. A szórendi variábilitás következtében pedig a pozíciós szókapcsolatok helyett relációs szókapcsolatokon célszerű jelöltlistát definiálni, ehhez viszont függőségi viszonyokat is tartalmazó annotáció kell.[10]

A fentiek alapján az alábbi főbb lépéseket látjuk fontosnak egy magyar szövegen működő TSZK kinyerő módszer létrehozásában:

- az azonosítani kívánt TSZK altípus illetve jelenség pontos meghatározása
- nyelvtani jellemzők, viselkedés részletes feltárása
- ennek alapján specifikus eljárás kidolgozása és alkalmazása.

Ezeket a lépéseket követő prototípus eljárást mutatunk be az 5. részben, előtte azonban röviden érintjük a kinyert és azonosított TSZK-k lexikonbeli reprezentációjának kérdését, majd összefoglaljuk azokat a nyelvi jelenségeket, amelyek diagnosztikai eszközként szolgálhatnak a kinyerési módszerek számára.

[9] Kis et al. (2004) ezt *típusos kollokációnak* nevezi.

[10] Amíg ez nem áll rendelkezésre, a POS annotáción működő reguláris szabályokkal közelíthető.

3.2. Reprezentáció

Ha rendelkezésünkre áll a TSZK-k listája és releváns tulajdonságaikat is meghatároztuk, a nyelvfeldolgozó rendszerekben történő hatékony hasznosíthatóságuk érdekében egyértelmű és gépileg kezelhető módon kell tárolnunk őket. A legegyszerűbb, változtathatatlan lexikai egységként (*word_with_spaces*), „listéma"-ként (Sciullo és Williams, 1987)való egyszerű felsorolás csak a *rögzített kifejezés* típusú TSZK-k kezelésére alkalmas. Egyéb esetben az összes variáns felsorolása kivihetetlen. A sztenderd nyelvtani szabályok által történő kezelés pedig a túlgenerálás és az idiomatikus jelentés származtatásának problémájával néz szembe.

Mivel jelenleg nincs kidolgozott, nyelvészetileg jól megalapozott, általánosan elfogadott számítógépes nyelvtani rendszer magyarban, jól kezelhető és a releváns tulajdonságokat egységesen leíró reprezentációs formalizmust pedig nehéz ettől teljesen függetlenül kidolgozni (Villavicencio et al., 2004), ismét egy opportunista következtetésre vagyunk kénytelenek jutni. A jelen helyzetben legfeljebb konkrét alkalmazásokhoz lehet specifikus TSZK (és nehézkesen hordozható) erőforrásokat fejleszteni.

4. TSZK-k nyelvi diagnosztikai eszközei

A következőkben néhány lehetséges eljárást vázolunk a TSZK-k automatikus kivonatolására. Ezek eltérő mértékben előfeldolgozott korpuszt igényelnek, ezért kívül eshetnek jelenlegi lehetőségeink határain — fontosságukat az adja, hogy kijelölnek néhány követhető jövőbeli kutatási irányt.

Elméleti szempontból a TSZK-k a „normálishoz képest" csökkent variabilitással bíró kifejezések. Bár ez a megfogalmazás a normalitás homályossága miatt maga is meglehetősen homályos, mégis talán úgy pontosítható, hogy a TSZK-k bizonyos, a szintaxis elvei által jósolt változatokkal nem rendelkeznek. Ez a hiány nem vezethető le sem a TSZK-ban előforduló kifejezések kategóriája, sem pedig a TSZK szintaktikai szerkezete alapján. Más szóval, az idiómák meglehetősen gyengén vethetők alá különböző szintaktikai transzformációknak idiomatikus jelentésük elvesztése nélkül. Ebben azonban az egyes idiómák között jelentős eltéréseket találunk. Az alábbiakban néhány példát mutatunk arra, ahol ez a jelenség tetten érhető.[11]

Eldöntendő érdés: *A bolondját járatod velem?* DE: **Hiszi a piszi?*

Kiegészítendő-kérdés: *Miféle vészharangot kongattak meg a tudósok?* DE: **Mit járattál Jánossal?*, **Miféle csatabárdot ástak ki Mariék?*, **Melyik kesztyűt vette fel Béla?*

Igeidő: *minden követ meg fog mozgatni, minden követ megmozgatott*, DE: **hinni fogja a piszi, *hitte a piszi.*

Progresszív (aspektus): *Amikor beléptem, János éppen húzta a lóbőrt*, DE: **Amikor beléptem, Jánosék éppen ásták ki a csatabárdot.* Ennek a magyarázata az lehet, hogy a *kiássa a csatabárdot* idiomatikus jelentése ('nyílt ellenségeskedésbe kezd') nem jól progresszivizálható (mivel egyfajta achievement), ezért az

[11] A '*', mint már említettük, itt az idiomatikus jelentés hiányát jelzi.

idióma formai progresszivizálása csak a szószerinti értelmet adhatja (hogy ui. Jánosék nagyban egy konkrét indián harci eszköz földből való kiemelésén dolgoztak).

Topikalizáció: *a rizsát, azt tudja nyomni, bakot, azt nem lőtt, de majdnem*, DE: **János a törülközőt bedobta, *a csatabárdot, azt kiásták/?a csatabárdot kiásták, *a kulcsot beadta a beteg, *a hajó elment, *a kesztyűt fel szokta venni.*

Fókusz: *'elásták a csatabárdot (és nem 'kiásták azt)*, DE: **a 'csatabárdot ásták el (és nem mást).*

Belső módosítás: *szép nagy bakot lőttél ezzel*, DE: **elment az utolsó hajó* (≠ 'elmulasztottad az utolsó lehetőséget').

Mellékmondatos módosíthatóság: *A vészharang, amit a tudósok megkongattak végül a washingtoni bürokratákat is felébresztette.* DE: **a bak, amit lőtt, végül az állásába került.*

Melléknévi igenévképzés: *a Jánossal a bolondját járató fiú, a csatabárdot újra kiásó ellenfelek, a minden követ megmozgató alperes*, DE: **az elmenő hajó* (≠ 'az eltűnő lehetőség'), *?a törülközőt bedobó ügyfél, *a kulcsot beadó beteg.*

Nominalizáció: általában nem lehetséges — **a jég (köztük történő) megtörése, *Jánosnak a Mari általi bolondját járatása, *a hajó elmenése, *a bak (le)lövése, ?a csatabárd kiásása, *minden kő megmozgatása*, bár pl. *a bili kiborulása nagy felbolydulást okozott* esetleg elfogadható idiomatikus jelentésben is.

A fenti tulajdonságok azonban — bár elméletileg relevánsak — közvetlenül nem használhatók fel jelenlegi céljaink eléréséhez, hiszen azt jelölik ki, ami nem létezik, míg a korpusz annak a tárháza, ami valóban aktualizálódott. Ezért az alábbiakban a TSZK-k olyan általános jellemzőire koncentrálunk, amelyek segítségével a korpuszból ténylegesen kinyerhetőkké válnak az ilyen kifejezések.

4.1. Tematikus inkongruencia

A TSZK-k egyik legalapvetőbb jellegzetessége (formai kötöttségük mellett), hogy jelentésük nem (teljesen) kompozícionális. Ugyanakkor azonban a TSZK-k igen nagy részéhez tartozik közvetlen, kompozícionális jelentés is. Ennek a ténynek a kommunikáció során is komoly jelentősége van, amelyet a kommunikáló partnerek kénytelenek figyelembe venni: *egy TSZK csak akkor használható idiomatikus jelentésében, ha kellő tematikus inkongruencia áll fenn a diskurzus tematikája és a TSZK kompozícionális jelentése között.* Az idiomatikus jelentés a szövegtematikába pragmatikai vagy egyéb okonál fogva be nem illeszthető kompozícionális jelentés kizárása után válik relevánssá a hallgató számára mint olyan „másodrendű jelentés," amelynek invokálásával képes elkerülni a kommunikáció megakadását (*Principle of Charity*, ld. Davidson (2001)). Ennek a körülménynek a számítógépes használatba vétele a szöveg „szemantikai súlypontjának" meghatározását igényli, ám ez elvileg és gyakorlatilag is lehetséges a jelenleg is használatban lévő vektoralapú szövegosztályozó eljárások segítségével. Az inkongruens, azaz nagy valószínűséggel idiomatikus jelentésű kifejezések hatása abban nyilvánulhat meg, hogy a szöveghez rendelt vektor a kifejezés hozzáadása után olyan mértékben megváltozik, amely egyébként a diszkurzus befejezését és egy új topik megnyitását jellemezné.

Bár a tematikus inkongruenciára építő eljárások teljes általánosságukban a diskurzustopik azonosítására alkalmas eszközöket igényelnek, az alábbi pontban tárgyalandó sajátos eset már szerényebb keretek között is detektálható lehet.

4.2. Szemantikai inkongruencia

A TKSZ-ek esetében sokszor találkozunk szemantikai furcsaságokkal: pl. *a szőnyeg alá söpri a problémát* esetében látszólag egy fizikai cselekvést (*seprés*) alkalmazunk egy absztrakt entitásra (*probléma*), ami első látásra kategóriahiba. Ehhez hasonló még: *húzza az időt, bedob egy új témát, kikerüli a választ.* Ez az inkongruencia már akkor is észlelhető, ha a korpusz szemantikailag legalábbis minimálisan annotálva van. Az ilyen szemantikailag anomális idiómák egyébként szintaktikailag úgy tűnik nagyobb variációs szabadságot is engednek meg: topikalizálhatók és belsőleg módosíthatók (*Az ilyen nem életbevágó problémát általában megpróbálják a szőnyeg alá söpörni az illetékesek, hacsak valaki meg nem akadályozza őket ebben*), továbbbá nominalizálhatóak is (*A probléma szőnyeg alá söprése nem jelent hosszú távú megoldást*). Ez valószínűleg azzal függ össze, hogy ez a fajta nyilvánvaló szemantikai anomália önmagában is elegendő az idiomatikus jelentés detektálásához minden környezetben, ezért nincs szükség további megszorítások kirovására.

4.3. Lexikai idioszinkrázia

Láttuk, hogy a TSZK-k alkatrészei sokkal kisebb variálhatósággal bírnak, mint a a teljesen produktív kifejezéseké. Ez többek között abban is megnyilvánul, hogy a TSZK részei többnyire nem cserélhetők fel (közel) szinonim kifejezésekkel az idiomatikus jelentés elveszése nélkül (aminek következtében — ha történetesen csak az idiomatikus jelentés létezett — a kifejezés értelmetlenné is válhat). Pl.: *a* ^{OK}*bolondját/*hülyéjét/ *gyengeelméjűjét/*retardáltját járatja vkivel*, ill. *játssza az* ^{OK}*eszét/*értelmét/ *intellektusát.* Ezt a jelenséget nevezhetjük **lexikai idioszinkráziának** (a szemantikai ekvivalensek közül is csak a „megfelelő" szó illeszthető be). Feltételezhetjük, hogy a teljesen produktív szókapcsolatok jobban tűrik ezt a fajta behelyettesítést, így ennek vizsgálatával lehetőség nyílhat TSZK-k azonosítására, illetve a TSZK-k teljesen produktív szókapcsolatoktól történő elkülönítésére.

A jelen vizsgálatban a TSZK-knak ezt a vonását kíséreltük meg tesztelni egy gépi szinonímaszótár felhasználásával.

5. TSZK-k és produktív szókombinációk elválasztása

A szinonimahalmazokon belül a helyettesíthetőség mértéke gyakorlatilag bármilyen AM segítségével kifejezhető, hiszen a csupán a hasonló jelentésű elemek előfordulására vonatkozó vizsgálattal nem teszünk mást, mint az AM számára az eseményteret leszűkítjük, ezáltal a mérték pontosságát illetve zajmentességét

próbáljuk növelni.[12] Kézenfekvő választás mint AM az a *kölcsönös információ* (KI), melyet valamilyen formában a hasonló, szemantikai alapú behelyettesíthetőséget vizsgáló eljárásokban többen alkalmaztak. Lin (1999) függőségileg elemzett korpuszból kinyert ⟨*fej,relációtípus,módosító*⟩ hármasokat vizsgált a kompozícionalitás szempontjából, míg McCarthy et al. (2003) frazális igék kompozícionalitásának erősségére kapott mértéket vetette össze emberi megítéléssel, mindegyikük elég visszafogott eredménnyel. Ez is jelzi azt, hogy a helyettesíthetőségen alapuló tesztek nem nagyon alkalmasak a kompozícionalitás mértékének meghatározására (Baldwin et al., 2003), úgyhogy célszerű inkább azt feltételezni, hogy csupán a produktív kombinációkat képesek elválasztani a TSZK-któl általában. Ezért az alábbiakban mi is ezzel a feladattal próbálkozunk.

Első lépésben AM-ként a Pearce (2002) által javasolt, a KI értékhez hasonló „standardizált eltérést" használjuk az alábbi módon. A szinonimahalmazok forrásaként a Magyar Szókincstár (Kiss, 2001) elektronikus változata szolgál.[13] Jelöljük $\mathcal{D}$-vel a szótári adatbázist, ebből rendelhetjük hozzá egy-egy szóhoz a hozzá tartozó szinonimahalmazokat:

$$\mathcal{D} = \{S_1, S_2, S_3 \ldots\} \tag{1}$$

Egy F fogalom egy lehetséges lexikális megvalósítása legyen K többszavas kifejezés, melyre: $K = \langle w_1 \ldots w_n \rangle$, mint elemi esemény. Az F fogalom összes lehetséges megvalósítása alkotja az eseményteret, mely a következőképpen definiálható a szinonimahalmazokon:

$$\Omega(K) = \{w'_{1,n} : w'_i \in S_i, 1 \leq i \leq n\} \tag{2}$$

ahol S_i a w_i szónak megfeleltethető szinonimahalmaz.

Ha feltesszük, hogy K teljesen produktív kifejezés, és elemeit egymástól függetlenül választjuk a megfelelő szinonimahalmazból, akkor K_i előfordulási valószínűségét a következőképpen közelíthetjük:

$$\hat{p}(K_i) = \prod_{i=1}^{n} p_i(w_i) \tag{3}$$

$p_i(w_i)$ annak valószínűsége, hogy az S_i szinonimahalmazból éppen w_i-t választjuk:

$$p_i(w_i) = \frac{f(w_i|S_i)}{\sum_{w \in S_i} f(w|S_i)} \tag{4}$$

Az így kapott értéket összevethetjük az adott kifejezés tényleges előfordulásával:

$$p(K_i) = \frac{f(K_i)}{\sum_{K \in \Omega(K)} f(K)} \tag{5}$$

[12] Nyilván a szinonimahalmazok megkonstruálásához szükséges erőforrással ennek „meg is fizetjük az árát".

[13] Gépi úton korpuszból származtatott tezaurusz is használható (pl. Lin (1998)), ennek magyarra történő megalkotása és a két módszer alapján történő összehasonlítás azonban egy későbbi kutatás tárgya lehet. Pearce (2002) szintén kész adatbázist, WordNet synseteket alkalmaz.

A két érték közötti különbség, illetve ennek z-transzformáltja jelzi, hogy az adott kifejezés mennyire „tűri" a helyettesíthetőséget:

$$z_i = \frac{d_i}{\sigma(d)}, \quad \text{ahol} \quad d_i = p(K_i) - \hat{p}(K_i) \tag{6}$$[14]

Ha magas z értéket kapunk, a kifejezés TSZK-nak lenne tekinthető.

5.1. Kiértékelés

A vizsgálathoz szükséges jelöltlista előállításához az MNSZ (Váradi, 2002) (POS egyértelműsített[15]) teljes anyagát használtuk (153 millió szó), amelyből 3-féle listát állítottunk elő:

1. szomszédos melléknév+főnév (L1)
2. egy mondaton belüli igék és tárgyesetű főnevek minden lehetséges ige+főnév kombinációja (L2)
3. egy mondaton belüli határozott ragozású ige + tárgyesetű főnév párok (L3)[16].

Mivel ezúttal a kifejezések morfológiai idioszinkráziáját nem kívántuk vizsgálni, lemmatizált alakokkal dolgoztunk, és csak az 5-nél gyakrabban előforduló kombinációkat vettük figyelembe. Az értékelés során a fenti z értéken alapuló modellre M_z-vel hivatkozunk.

A kapott eredményeket egy olyan viszonyító alapmodellel vetettük össze, amelyben KI-t, illetve ennek t próbából számított küszöbbel szűrt változatát (Church et al., 1994) használtuk a teljes eseménytéren[17] (M_{va}). Az összehasonlítást érdemes elvégezni azzal a modellel is, amelyben a jelöltek rangsorolását a kifejezés elemeihez tartozó szinonimahalmazokból képezhető leggyakrabban előforduló (f' számú) elempár és a második leggyakoribb (f'' számú) elempár előfordulásából számított egyszerű gyakorisági arány ($s = \frac{f'-f''}{f'}$) végzi (M_s).

Az értékelésben gyakorlati szempontok miatt a legelterjedtebb, *legjobb n-lista* módszert alkalmaztuk: az egyes modellek által rangsorolt jelöltlistákból kiválasztottuk az első n (itt $n = 250$) kifejezést, és megnéztük, milyen arányban tartalmaznak TSZK-nak tekinthető szókapcsolatot (*pontosság*). Ugyan sokan (pl. Evert és Krenn (2001)) számolnak a másik közkeletű mérőszámmal is (*fedés*), egyszerűen belátható, hogy az ilyen típusú kiértékelési feladatokban ez semmiféle további új információt nem ad a felhasznált modellek minőségével kapcsolatban a *pontossághoz* képest[18]. Emiatt az 1. táblázat csupán *pontosság* értékeket tartalmaz.

[14] $\sigma(d) = \sqrt{\frac{\sum_i (d_i - \mu)^2}{n}}$

[15] Az egyértelműsítés hibája kb. 3%, ez a listában elkerülhetetlen zajhoz vezet.

[16] Szintaktikai annotáció hiányában ezzel az egyszerű heurisztikával próbáltunk közelíteni valamiféle fej-argumentum viszonyt.

[17] Vagyis azt mérjük, hogy a kifejezés mennyire összetartozó, de nem a szinonimahalmazokból alkotható variánsokhoz, hanem az összes lehetséges jelöltkombinációhoz képest.

[18] A *legjobb n-lista* esetén ugyanis a *pontosság*: $p(n) = \frac{TP(n)}{n}$, ahol TP(n) az n elemet tartalmazó listában a helyes találatok száma (*true positive*). A fedés ugya-

1. táblázat. A modellek teljesítménye.

Lista	Jelöltek száma	Pontosság $p(n = 250)$		
	$f(K_i) > 5$	M_{va}	M_s	M_z
L1	191454	54.4%	15.2%	17.2%
L2	452981	29.2%	4.8%	19.2%
L3	100559	56.8%	9.2%	38.0%

Az alapmodell (M_{va}) hasonló eredményt ad, mint a szakirodalomban található *pontosság* értékek, és igazolja a szokásos elvárást is: minél megszorítottabb a jelöltlista, annál hatékonyabb a kinyerő eljárás. A szinonimahalmazokban való helyettesíthetőségen alapuló modellek kudarcának oka véleményünk szerint az, hogy a 4.3. részben megfogalmazott, a *lexikai idioszinkráziát* kihasználó, és sokak (Lin, 1999; Pearce, 2001; McCarthy et al., 2003; Bannard et al., 2003) által használt hipotézis ugyan lehet, hogy igaz, de ez a korpuszban nem érhető tetten: nincs olyan mértékű mérhető különbség a teljesen produktív kifejezések elemeinek helyettesíthetőségében egy TSZK-k elemeihez képest, amit egy azonosító módszer jól fel tudna használni. Nem állítjuk azonban, hogy az eféle módszer teljesen haszontalan; előfordulhat, hogy egy nagyon részletesen specifikált TSZK altípus kinyerésében mégis használható[19], mindazonáltal egy általában szokásos módon megszorított jelöltlista esetén nem hoz eredményt.

6. Összefoglalás és további feladatok

A dolgozatban megkíséreltünk áttekintést adni a többszavas kifejezések számítógépes kezelése során felmerülő kérdésekről, és megoldandó feladatokról. Megpróbáltuk összefoglalni azokat a nyelvi jelenségeket, amelyek segítségével azonosító, kinyerő eljárások építhetőek, és megmutattuk, hogy a hasonló jelentésű elemek helyettesíthetőségén alapuló eljárások nem állnak szilárd, nagy korpuszból nyert adatokkal alátámasztható alapon.

Természetes további feladatként adódik magyar nyelven olyan vizsgálatok elvégzése, amelyek további információt adnak kinyerési módszerek alkalmazhatóságáról és hatékonyságáról: több féle AM alkalmazása és összehasonlító kiértékelése, illetve adott TSZK-típus azonosításához a legalkalmasabb AM kiválasz-

nitt: $f(n) = \frac{TP(n)}{C}$, ahol C a teljes listában található TP-k száma, ami konstans. Ekkor $f(n) = \frac{p(n) \times n}{C}$, vagyis a *fedés* valójában a *pontosság* információmentes, modellfüggetlen transzformáltja. (Ha $n = |$*teljes jelöltlista*$|$, akkor természetesen $f(n) = \frac{\frac{C}{n} \times n}{C} = 1 \rightarrow 100\%$. A *fedés* legfeljebb annyiban lehet informatív, hogy az adott módszerhez hozzá lehet rendelni azt a legkisebb $n < C$ értéket, melyre $f(n) = 1$, vagyis azt a minimális listanagyságot, amelyben az összes TP már benne van. Ezt a *pontosság* függvényről nem lehet leolvasni.)

[19] Ennek vizsgálata további feladat.

tása és kidolgozása. Ez a munka, konkrétan például a morfológiai idioszinkrázia jelenségét kihasználó módszer fejlesztése, jelenleg is folyik.

Hivatkozások

Baldwin, Timothy, Bannard, Colin, Tanaka, Takaaki és Widdows, Dominic. An Empirical Model of Multiword Expression Decomposability. In: *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan. 2003, 89–96.

Bannard, Colin, Baldwin, Timothy és Lascarides, Alex. A Statistical Approach to the Semantics of Verb-Particles. In: *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan. 2003.

Bauer, Laurie. *English Word-Formation.* Cambridge University Press, Cambridge, England, 1983.

Bourigault, Didier. Lexter, a Natural Language Processing Tool for Terminology Extraction. In: *Proceedings of 7th EURALEX International Congress*, 1996.

Calzolari, Nicoletta, Fillmore, Charles J., Grishman, Ralph, Ide, Nancy, Lenci, Alessandro, MacLeod, Catherine és Zampolli, Antonio. Towards Best Practice for Multiword Expressions in Computational Lexicons. In: *Poceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas, Spain. 2002, 1934–40.

Chomsky, Noam. *Rules and Representations.* Columbia Univeristy Press, New York, 1980.

Church, Kenneth W. és Hanks, Patrick. Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics*, 1990, 16(1):23–29.

Church, Kenneth Ward, Gale, William, Hanks, Patrick, Hindle, Donald és Moon, Rosamund. Lexical Substitutability. In: Atkins, B. T. S. és Zampolli, Antonio szerk. *Computational Approaches to the Lexicon.* Oxford University Press, 1994, 153–180.

Daille, Béatrice. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In: Klavans, Judith és Resnik, Philip szerk. *The Balancing Act: Combining Symbolic and Statistical Approaches to Language.* The MIT Press, Cambridge, Massachusetts, 1996, 49–66.

Davidson, Donald. Radical Interpretation. In: *Inquiries into Truth and Interpretation.* Clarendon Press, Oxford, 2001.

Dunning, Ted. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 1993, 19(1):61–74.

Evert, Stefan és Krenn, Brigitte. Methods for the qualitative evaluation of lexical association measures. In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France. 2001, 188–195.

Forgách, Tamás. Frazeológia és valencia. In: Büky, László és Maleczki, Márta szerk. *A mai magyar nyelv leírásának újabb módszerei*, III, 1998, 7–39.

Heid, Ulrich. Extracting Terminologically Relevant Collocations from German Technical Texts. In: Sandrini, P. szerk. *TKE99 Terminology and Knowledge*

Engineering. Proceedings Fifth International Congress on Terminology and Knowledge Engineering, Vienna. 1999, 241–255.

Kaalep, Heiki-Jaan és Muischnek, Kadri. Using the Text Corpus to Create a Comprehensive List of Phrasal Verbs. In: *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas, Spain. 2002, 101–105.

Kis, Balázs és Ugray, Gábor. Új korpuszstatisztikai eszköztár kollokációkeresésre. In: *Magyar Számítógépes Konferencia*, Szeged. 2003, 131–136.

Kis, Balázs, Villada, Begoña, Bouma, Gosse, Ugray, Gábor, Bíró, Tamás, Pohl, Gábor és Nerbonne, John. A New Approach to the Corpus-based Statistical Investigation of Hungarian Multi-word Lexemes. In: *Poceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisbon, Portugal. 2004, 1677–1680.

Kiss, Gábor szerk. *Magyar Szókincstár*. Tinta Könyvkiadó, Budapest, 2001.

Krenn, Brigitte és Evert, Stefan. Can we do better than frequency? A case study on extracting PP-verb collocations. In: *Proceedings of the ACL Workshop on Collocations*, Toulouse, France. 2001, 39–46.

Lin, Dekang. Automatic retrieval and clustering of similar words. In: *Proceedings of COLING/ACL-98*, Montreal. 1998, 768–774.

Lin, Dekang. Automatic identification of noncompositional phrases. In: *Proceedings of the 37th Annual Meeting of the ACL*, College Park, USA. 1999, 317–24.

Manning, Christopher és Schütze, Hinrich. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA, 1999.

McCarthy, Diana, Keller, Bill és Carroll, John. Detecting a Continuum of Compositionality in Phrasal Verbs. In: *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan. 2003.

McKeown, Kathleen R. és Radev, Dragomir R. Collocations. In: Dale, Robert, Moisl, Hermann és Somers, Harold szerk. *A Handbook of Natural Language Processing*. Marcel Dekker, 2000.

Nunberg, Geoffrey, Sag, Ivan A. és Wasow, Thomas. Idioms. *Language*, 1994, 70(3):491–538.

Oflazer, Kemal, Çetinoğlu, Özlem és Say, Bilge. Integrating Morphology with Multi-word Expression Processing in Turkish. In: Tanaka, Takaaki, Villavicencio, Aline, Bond, Francis és Korhonen, Anna szerk. *Second ACL Workshop on Multiword Expressions: Integrating Processing*, Barcelona, Spain. Association for Computational Linguistics, July, 2004, 64–71.

Pearce, Darren. Using conceptual similarity for collocation extraction. In: *Proceedings of the 4th UK Special Interest Group for Computational Linguistics*, 2001.

Pearce, Darren. A comparative evaluation of collocation extraction techniques. In: *Poceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas, Spain. 2002.

Pulman, Stephen G. The recognition and interpretation of idioms. In: Cacciari, Cristina és Tabossi, Patrizia szerk. *Idioms: Processing, Strucutre and Interpretation*, 11. fejezet. Lawrence Erlbaum Associates, Hillsdale, NJ, 1993.

Sag, Ivan, Baldwin, Timothy, Bond, Francis, Copestake, Ann és Flickinger, Dan. Multiword Expressions: A Pain in the Neck for NLP. In: *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, Mexico City, Mexico. 2002, 1–15.

Schone, Patrick és Jurafsky, Daniel. Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords A Solved Problem? In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Pittsburgh, PA. 2001.

Sciullo, Anna-Maria Di és Williams, Edwin. *On the Definition of Word.* MIT Press, Cambridge, MA, 1987.

Shimohata, Sayori, Sugio, Toshiyuki és Nagata, Junji. Retrieving Collocations by Co-occurrences and Word Order Constraints. In: *Proceedings of ACL-EACL 97*, 1997, 476–481.

Villavicencio, Aline, Copestake, Ann, Waldron, Benjamin és Lambeau, Fabre. Lexical Encoding of MWEs. In: Tanaka, Takaaki, Villavicencio, Aline, Bond, Francis és Korhonen, Anna szerk. *Second ACL Workshop on Multiword Expressions: Integrating Processing*, Barcelona, Spain. Association for Computational Linguistics, July, 2004, 80–87.

Váradi, Tamás. The Hungarian National Corpus. In: *Proceedings of the Second International Conference on Language Resources and Evaluation*, Las Palmas. 2002, 385–389.